RESEARCH ARTICLE

PlasmoSEP: Predicting surface-exposed proteins on the malaria parasite using semisupervised self-training and expert-annotated data

*Yasser El-Manzalawy[1], Elyse E. Munoz[2], Scott E. Lindner[2] and Vasant Honavar[1]**

[1] College of Information Sciences and Technology, Pennsylvania State University, PA, USA
[2] Center for Malaria Research, Department of Biochemistry and Molecular Biology, Pennsylvania State University, PA, USA

Accurate and comprehensive identification of surface-exposed proteins (SEPs) in parasites is a key step in developing novel subunit vaccines. However, the reliability of MS-based high-throughput methods for proteome-wide mapping of SEPs continues to be limited due to high rates of false positives (i.e., proteins mistakenly identified as surface exposed) as well as false negatives (i.e., SEPs not detected due to low expression or other technical limitations). We propose a framework called PlasmoSEP for the reliable identification of SEPs using a novel semisupervised learning algorithm that combines SEPs identified by high-throughput experiments and expert annotation of high-throughput data to augment labeled data for training a predictive model. Our experiments using high-throughput data from the *Plasmodium falciparum* surface-exposed proteome provide several novel high-confidence predictions of SEPs in *P. falciparum* and also confirm expert annotations for several others. Furthermore, PlasmoSEP predicts that 25 of 37 experimentally identified SEPs in *Plasmodium yoelii* salivary gland sporozoites are likely to be SEPs. Finally, PlasmoSEP predicts several novel SEPs in *P. yoelii* and *Plasmodium vivax* malaria parasites that can be validated for further vaccine studies. Our computational framework can be easily adapted to improve the interpretation of data from high-throughput studies.

Received: June 8, 2016
Revised: August 31, 2016
Accepted: October 5, 2016

Keywords:
Bioinformatics / Malaria / *Plasmodium* / Predicting surface-exposed proteins / Semi-supervised learning / Surface-exposed proteomics

Additional supporting information may be found in the online version of this article at the publisher's web-site

1 Introduction

Malaria remains one of the largest global health burdens today, with an estimated 438 000 deaths and 214 million new infectious occurring annually [1]. This disease is caused by a eukaryotic parasite of the genus *Plasmodium* that is transmitted by infected *Anopheles* mosquitoes. Five *Plasmodium* species infect humans, including *P. falciparum* and *P. vivax*, which together cause nearly all of the mortalities and morbidities. In addition, there are several *Plasmodium* species that infect small animals, and thus serve as excellent models of infection (e.g. *P. yoelii* and *P. berghei* in mice, *P. cynomolgi* in nonhuman primates). These parasites (except *P. vivax*, which cannot be continuously passaged in the laboratory) have been used to identify weaknesses in the parasite that can be exploited for chemotherapies and vaccine candidates.

Correspondence: Dr. Scott E. Lindner, W223 Millennium Science Complex, University Park, PA 16802, USA
E-mail: Scott.Lindner@psu.edu

Abbreviations: **AUC**, area under ROC curve; **IMC**, inner membrane complex; **NB**, Naïve Bayes; **RF**, random forest; **SEPs**, surface-exposed proteins; **SL**, supervised learning; **SSL**, semi-supervised learning

*Additional corresponding author: Professor Vasant Honavar
E-mail: vhonavar@ist.psu.edu

Significance of the study

Profiling the surface exposed proteome of the malaria parasite is of major importance for understanding host-parasite interactions and for identifying novel subunit vaccine candidates. MS-based proteomic techniques have increasingly become the state-of-the-art experimental approach for mapping surface exposed proteins (SEPs) in many target pathogens. However, more efforts are needed to improve the reliability of the interpretation of the results of such experiments. We propose a novel computational approach to effectively postprocess MS results and filter out false positive as well as false-negative results. Specifically, we integrate imperfect results of MS experiments for mapping SEPs in *P. falciparum*, expert annotation of these data, and semi-supervised machine learning approaches to develop prediction models that could be used to: (i) validate the output of MS experiments; (ii) predict novel SEPs that have been missed by the MS experiments; (iii) predict novel SEPs in different species of *Plasmodium* (e.g., *P. yoelii* and *P. vivax*). This study, which, to the best of our knowledge, is the first study of its kind, opens up opportunities for developing community resources that integrate and improve the reliability of SEPs identified by high-throughput MS-based proteomic experiments.

Current efforts to reduce and eliminate parasite transmission have relied upon controlling the mosquito vector, supplying insecticide-treated bednets, and administering antimalarial drugs that kill the blood stage of the parasite. In contrast to these efforts, the development of an effective vaccine against *P. falciparum* and *P. vivax* has encountered several barriers, and to date no licensed vaccine candidate has reached the levels of protection thought to be required to make a substantial impact upon parasite transmission (reviewed in [2]). The most advanced vaccine candidate (called RTS,S) provides limited, short lived protection in Phase III clinical trials in Africa, but has served as an important first milestone [3]. The RTS,S vaccine consists of a single surface protein (circumsporozoite protein, CSP) that is present on the sporozoite form of the parasite, which is transmitted from mosquitoes to humans. As CSP is known to have considerable variation in field isolates (*ibid*), parasites are likely to evade antibody-based immune responses by simply changing the composition of this protein. Ongoing efforts now aim to improve upon RTS,S by adding additional antigens to create bivalent or multivalent vaccine candidates, and by using alternate delivery approaches (e.g. viral vectors) [4]. However, the experimental validation of surface-exposed proteins (SEPs) that will be accessible to antibodies has been limited in scope, and thus the list of vetted antigens available for multivalent vaccines has remained short.

Our previous work has provided an initial, and then more recently a comprehensive, list of proteins on the surface of the transmitted sporozoite form of the parasite, which are accessible to antibody-based immune responses [5, 6]. Taken together, these catalogues of SEPs provide a much needed, experimentally validated list to draw upon to design next generation, multivalent malaria vaccines. These studies have focused primarily upon *P. falciparum*, which can be grown in the laboratory and is thus amenable to these studies.

In the absence of data describing the SEPs in other human-infectious malaria parasites, it would be advantageous to draw upon our current knowledge of the surface proteome of sporozoites to accurately predict which proteins may be targetable. The supervised machine learning approach [7] is an efficient and cost-effective approach to extract hidden patterns from data (e.g., *P. falciparum* SEPs) and train predictive models that could be applied to predict novel SEPs in other human-infectious malaria parasites. However, the reliability of the predictions depends mainly on the quality of the training data. Taking into account the technical limitations of MS techniques [8, 9], our identified *P. falciparum* SEPs are expected to have a significant number of false positives (labeled cytosolic proteins from dying cells) as well as false negatives (due to limits of detection and sample scarcity), making the applicability of supervised machine learning algorithms to learn from such data a practical challenge. To address this challenge, we propose a novel framework for developing reliable predictive models from noisy high-throughput *P. falciparum* surface exposed proteomic data. Our approach integrates expert annotation of high-throughput data and semisupervised machine learning algorithms [10] to develop reliable predictive models for predicting SEPs in *Plasmodium*. Our results using simulated datasets acknowledge the viability of semi-supervised learning (SSL) to develop classifiers from small-size labeled data by exploiting available unlabeled data. Moreover, we demonstrate improvements in performance of SSL by leveraging noisy expert-annotated data. Finally, we have extended our approach to predict SEPs in human-infectious *P. vivax*, which cannot be continuously cultured in the laboratory. Taken together, here we provide the scientific community with predicted and experimentally validated SEPs for different *Plasmodium* species, along with an algorithm to help guide experimental validations of these proteins' potential as malaria subunit vaccine candidates.

2 Materials and methods

2.1 Surface-exposed proteomics

The surface exposed proteome of *P. yoelii* sporozoites was determined as previously described with few modifications

[5,6]. The detailed procedure is reported in Supporting Information Materials and Methods.

2.2 SSL

SSL [10, 11] is a class of supervised learning (SL) that makes use of available (often large amounts of) unlabeled data to train a model using a small set of costly labeled data. Many machine-learning researchers reported considerable improvements in classifier performance when unlabeled data is used in conjunction with small-size training data as opposed to building the model using only available labeled training data. In Supporting Information Materials and Methods, we summarize the self-training algorithm [10], a commonly used semisupervised algorithm that has been successfully used for various SSL tasks in Bioinformatics applications (e.g., [12–17]).

2.3 Self-training with potentially labeled data

In some applications, in addition to the labeled data L and unlabeled data U, we may have access to potentially labeled data P where the labels are based on information (e.g., expert opinion) that may be less reliable than direct experimental evidence. In Supporting Information Materials and Methods, we present a natural extension of the self-training algorithm to SSL that takes advantage of such potentially labeled data when available. Our Java implementation of the proposed self-training algorithm has been made freely available to the broader research community as part of the EpiT tool [18] (available at http://ailab.ist.psu.edu/epit/). This allows our algorithm to be invoked using the WEKA GUI [19] and to take advantage of several amino acid sequence derived features for building classifiers using EpiT.

2.4 Our framework

Figure 1 summarizes our proposed framework for improving the reliability of the results of high-throughput proteomics experiments for identifying SEPs. First, the output of one or more high-throughput proteomics studies for identifying SEPs in *P. falciparum* is used to generate a set of non-SEPs from the entire *P. falciparum* proteome (see Section 2.5 for details); Experimentally identified candidate SEPs are annotated by domain expert(s) as known SEPs, likely SEPs, unlikely SEPs, and unknown SEPs. Second, our novel self-training algorithm is applied to labeled, annotated, and unlabeled *P. falciparum* data to build a classifier for predicting SEPs in *P. falciparum,* PlasmoSEP. Third, the PlasmoSEP classifier and two additional Bioinformatics tools (SignalP [20] and an in-house model trained to predict protective antigens in parasites) are integrated together into a final prediction model that returns the maximum prediction score from these three

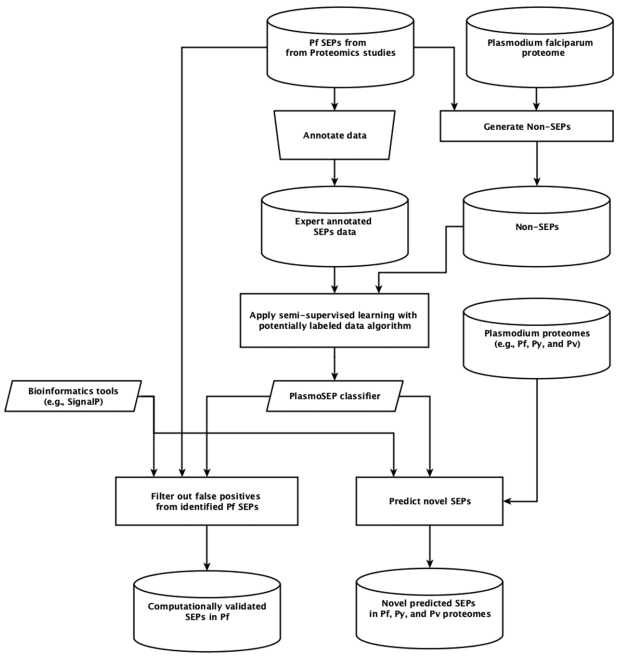

Figure 1. Flowchart of PlasmoSEP framework for integrating proteomics studies, expert annotations, bioinformatics tools, and semisupervised learning for accurate identification of SEPs in the malaria parasite (*Plasmodium* spp.).

predictors. The final model is then used to predict SEPs from among the experimentally identified candidate SEPs. This helps filter out false positives from proteomics experiments, which have been especially problematic in recent studies. Finally, an integrated model is also used to identify novel SEPs from entire *P. falciparum* proteomes and proteomes of other related malaria species, *P. yoelii and P. vivax*.

2.5 Classification experiments

We experimented with the two self-training algorithms using simulated and real-world datasets. Detailed description of the datasets, the extracted features, and the experimental settings are provided in Supporting Information Materials and Methods.

2.6 Other sources of information

To improve the reliability of our predicted SEPs, we used two additional types of evidence for complementing the predictions supplied by the PlasmoSEP classifier: (i) prediction of signal peptides provided by SignalP Web server [20] (as the presence of a signal peptide in a protein suggests that the protein is secreted or is a membrane protein [21, 22]); and (ii) prediction of protective antigens in parasites provided by our in-house classifier, described in Supporting Information Materials and Methods.

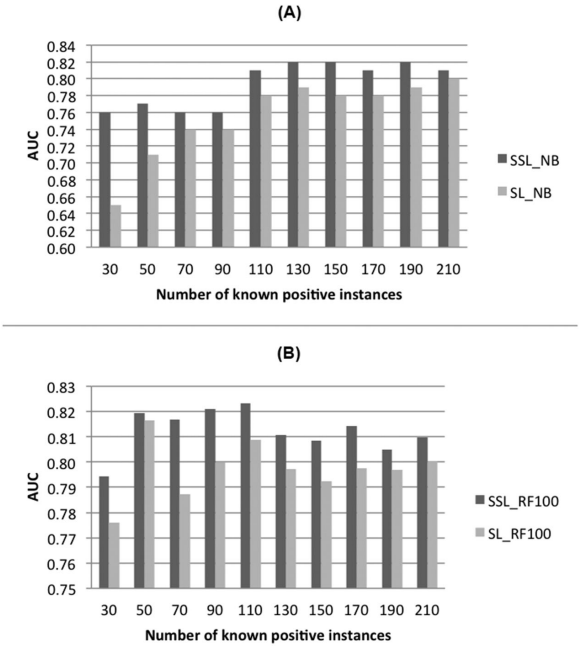

Figure 2. AUC comparisons between supervised learning (SL) and semisupervised learning (SSL), Algorithm 1, using NB (top) and RF100 (bottom) as supervised and base classifiers.

3 Results and discussion

3.1 Predictive models trained using semisupervised methods outperform those trained by their supervised counterparts

First, we compared the performance of SL and self-training (both using two basic SL algorithms, Naïve Bayes (NB) and random forest with 100 trees (RF100)) on the simulated datasets. In this experiment, the training dataset was randomly partitioned into labeled data $L = \{L^+ \cup L^- | \, s.t. \, |L^+| = |L^-|\}$ and unlabeled data U. Figure 2 reports the area under ROC curve (AUC) [23] estimated using the *independent* test data for NB (top) and RF100 (bottom) trained using only L and self-training classifiers using an NB and RF100 classifiers trained on L and U for different choices of $|L^+|$. Our results show that when the number of labeled data samples is small, the self-training algorithm substantially outperforms its supervised counterpart. For all choices of $|L^+|$ considered in this experiment, the classifiers trained using the semisupervised algorithm (SSL_NB and SSL_RF100) consistently outperform those trained using their supervised counterparts (SL_NB and SL_RF100).

3.2 Noisy expert-annotated data improve the performance of models trained using SSL

The results summarized in Fig. 2 suggest that when $|L^+|$ is less than 110 positive samples or, in other words, when the size of the labeled training data is less than 220 samples, the AUC of the SSL model is less than 0.80. To examine whether potentially labeled data (e.g., expert-annotated data) improve the performance of the semisupervised self-training algorithm, Algorithm 2, we designed the following experiment. We set $|L^+|$ to 90 and we randomly selected a subset of the remaining training data as potentially labeled such that potentially labeled data include an equal number of positively and negatively labeled samples (e.g., $|P^+| = |P^-|$). We also experimented with different numbers of potentially labeled data samples ($|P^+| = \{10, 30, 50, 70, 90\}$) and different levels of randomly added noise to the labels of the potentially labeled data. Using NB as the base classifier, we found that, for all choices of the number of potentially labeled samples and the fraction that are incorrectly labeled, predictive models trained using our proposed semisupervised self-training algorithm that takes advantage of potentially labeled data outperform those that do not (Fig. 3A). We repeated the experiment using RF100 as the base classifier for the two self-training algorithms (with $|L^+|$ set to 30 because for $|L^+|$ greater than 30, the AUC of SSL_RF100 is greater than 0.80 and very close to the upper limit for performance obtained using RF100 and the entire training dataset). Figure 3B shows that the proposed SSL algorithm using noisy potentially labeled data

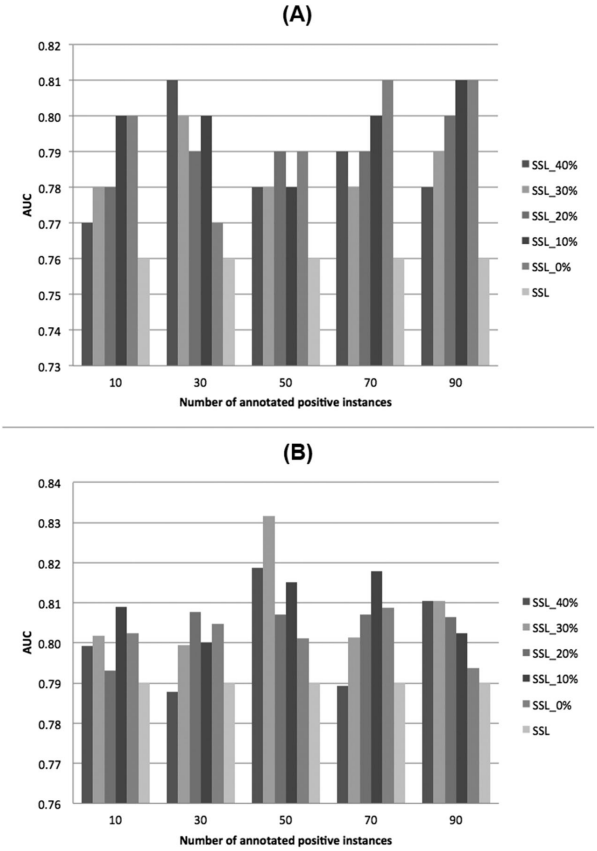

Figure 3. AUC comparisons between basic SSL and our proposed SSL (SSL_k%) with k% noise in potentially labeled data using NB (top) and RF100 (bottom) as base classifiers.

consistently outperforms the basic SSL algorithm so long as the fraction of potentially labeled data with incorrect labels does not exceed 30%. Interestingly, in some of the cases, the models trained on noisy potentially labeled data outperform or perform as well as those trained on accurate potentially labeled data. This observation may be explained in part by the theoretical results that suggest that noise in the training data mimics the behavior of regularization which in turn helps reduce overfitting and improve generalization, at least in the supervised setting [24]. This finding suggests that the proposed approach to SSL can take advantage of noisy potentially labeled data.

3.3 PlasmoSEP predicted SEPs in *P. falciparum*

We applied our proposed semisupervised self-training algorithm to learn from a *P. falciparum* surface-exposed proteome dataset, as described in Section 2. Briefly, each protein sequence is represented using its composition transition distribution features (see Supporting Information Materials and Methods) and we applied Algorithm 2 to train a RF classifier with 100 trees (RF100) from labeled, unlabeled, and potentially labeled data. Our choice of a RF100 classifier is based on its superior performance observed on the simulated data and its ability to cope with high-dimensional data [25]. The final learned model was then applied to the total *P. falciparum* salivary gland sporozoite proteome (2003 proteins) and a score was assigned to predict the probability that the protein is surface exposed. Two additional scores for each protein are obtained using SignalP and our in-house protective antigenicity predictor, respectively. Supporting Information Table 1 reports the three scores for every protein in the *P. falciparum* salivary gland sporozoite proteome and ranks all proteins using the maximum of the three scores. Supporting Information Table 2 reports the top 190 ranked proteins, those with a maximum score ≥ 0.7. Interestingly, this set of 190 predicted *P. falciparum* SEPs covers the 13 known SEPs (Supporting Information Table 3), 11 of 24 proteins annotated by expert curation as likely to be SEPs (Supporting Information Table 4), and only four of 41 proteins tagged as unlikely to be SEPs (Supporting Information Table 5). Finally, of 127 proteins tagged as unknown by expert curation, 23 proteins are predicted to be SEPs (Table 1). The top hit in the prediction for both *P. falciparum* and *P. yoelii* (see below), which also scored positively for *P. vivax* was GAPDH. Interestingly, while GAPDH is commonly regarded as a cytosolic housekeeping protein, it was recently shown experimentally to also be a bona fide surface antigen, and thus likely has moonlighting functions on the cell surface as well [26]. We hold that predicted SEPs in these lists, which may similarly be dismissed due to having a well-known/canonical cytosolic function, should be considered as having a possible surface function.

It should be noted that neither SignalP, a program for predicting secreted proteins, nor our in-house classifier for predicting protective antigens in parasites, on its own, is sufficiently reliable as a predictor of SEPs. However, because any protective antigen is essentially surface exposed or an exported protein [27, 28] and secreted proteins are frequently (but not always) retained on the cell surface [5, 6], we employ these two predictors to aid in the identification of potential SEPs that are not detected by our PlasmoSEP classifier. Therefore, the final PlasmoSEP score is set to be the maximum of scores predicted by the PlasmoSEP, SignalP, and the antigenicity classifiers. For example, if a query protein is assigned a low prediction score by PlasmoSEP classifier and a high score by SignalP and/or the antigenicity classifier, then we conclude that PlasmoSEP prediction is more likely to be a false negative and we return the high score assigned by SignalP and/or the antigenicity classifier as our final predicted score. On the other hand, if a query protein is assigned a high score by PlasmoSEP classifier but low scores by SignalP and/or the antigenicity classifier, then we conclude that the query protein is likely to be surface exposed that is not secreted or not a putative protective antigen.

Finally, we found that our predicted SEPs are consistent with their known biological roles in the parasite. Thus, the predicted SEPs include invasion-related proteins such as rhoptry neck proteins (ASP, RON2, RON3) and sporozoite invasion-associated protein 1 (SIAP1), surface adhesion proteins (CSP, TRAP), members of the gliding motility/inner membrane complex (IMC) apparatus, proteases (ROM4), perforins to aid cell traversal (PLP1), and metabolite transporters (Supporting Information Tables 2 and 3). As shown by previous experimental studies [6], even proteins that are transiently exposed to the surface during gliding, traversal, or invasion cues are truly surface exposed and are accessible to membrane impermeable labeling reagents and antibodies.

3.4 PlasmoSEP predicted SEPs in *P. yoelii* salivary gland sporozoites

In order to demonstrate the utility of the PlasmoSEP predictor across *Plasmodium* species, we next applied PlasmoSEP to the proteome of the rodent-infectious malaria species *P. yoelii*. Our previous studies, which first described an approach to identify the surface-exposed proteome of sporozoites, uncovered only a small number of proteins that were surface exposed on *P. yoelii* [5]. As this small number of proteins is insufficient for robustly testing the algorithm, we have built upon these initial findings and have now used improved labeling and washing conditions to expand the high confidence surface-exposed proteome (Table 2, Supporting Information Table 6). In brief, highly purified sporozoite samples were split just prior to addition of the biotin-conjugated crosslinker, with one half receiving the disulfide-containing labeling reagent (EZ-Link Sulfo-NHS-SS-Biotin) and the other half remaining unlabeled. These matched controls were otherwise treated identically throughout the experiment, including a high stringency washing protocol in urea and SDS. Together,

Table 1. List of predicted *P. falciparum* SEPs with maximum score ≥ 0.70 from the set of expert annotated unknown SEPs

ID	Name	PlasmoSEP	SignalP	Antigenicity	Max_score
PF3D7_1462800	Glyceraldehyde-3-phosphate dehydrogenase (GAPDH)	1.00	0.17	0.35	1.00
PF3D7_0818900	Heat shock protein 70 (HSP70)	1.00	0.11	0.56	1.00
PF3D7_1444800	Fructose-bisphosphate aldolase (FBPA)	1.00	0.10	0.25	1.00
PF3D7_0903700	Alpha tubulin 1	1.00	0.14	0.38	1.00
PF3D7_0922200	S-adenosylmethionine synthetase (SAMS)	1.00	0.11	0.32	1.00
PF3D7_0627500	4-Methyl-5(B-hydroxyethyl)-thiazol monophosphate biosynthesis enzyme	1.00	0.12	0.60	1.00
PF3D7_1140400	Conserved Plasmodium protein, unknown function	1.00	0.10	0.62	1.00
PF3D7_1133400	Apical membrane antigen 1 (AMA1)	0.00	0.55	0.91	0.91
PF3D7_1235700	ATP synthase subunit beta, mitochondrial	0.90	0.15	0.27	0.90
PF3D7_0826700	Receptor for activated c kinase (RACK)	0.90	0.10	0.41	0.90
PF3D7_0620000	Conserved Plasmodium protein, unknown function	0.00	0.87	0.58	0.87
PF3D7_1335900	Sporozoite surface protein 2 (TRAP)	0.10	0.85	0.30	0.85
PF3D7_1028600	Conserved Plasmodium protein, unknown function	0.00	0.10	0.85	0.85
PF3D7_0812300	Conserved Plasmodium protein, unknown function	0.00	0.84	0.39	0.84
PF3D7_0917900	Heat shock protein 70 (HSP70-2)	0.10	0.84	0.48	0.84
PF3D7_0513300	Purine nucleoside phosphorylase (PNP)	0.80	0.12	0.56	0.80
PF3D7_0524000	Karyopherin beta (KASbeta)	0.00	0.10	0.78	0.78
PF3D7_0708400	Heat shock protein 90 (HSP90)	0.10	0.13	0.75	0.75
PF3D7_0827900	Protein disulfide isomerase (PDI8)	0.00	0.74	0.35	0.74
PF3D7_1361800	Conserved Plasmodium protein, unknown function	0.00	0.11	0.70	0.70
PF3D7_0922500	Phosphoglycerate kinase (PGK)	0.00	0.10	0.70	0.70
PF3D7_0320300	T-complex protein 1 epsilon subunit, putative	0.00	0.10	0.70	0.70
PF3D7_1037300	ADP/ATP transporter on adenylate translocase	0.70	0.19	0.19	0.70

the high stringency washes and improved elution conditions (e.g. reducing the disulfide bond present in the crosslinker) reduced background contamination substantially, with only one and two proteins being captured in the unlabeled control replicates [5].

Several aspects of these data indicate that these proteins are bona fide surface proteins. As anticipated, many well-known SEPs are detected in this experimentally defined list, including CSP, TRAP, SPECT2, SIAP1, GAMA, TRSP, hexose transporter, and many others [5, 29–32]. Many of these proteins serve as the basis for existing malaria vaccine antigens (CSP, TRAP) or are the chosen targets for chemotherapeutics (hexose transporter) [2, 29]. Additionally, proteins involved in the IMC that are used by the invasive forms of the parasite for locomotion (termed gliding motility) were also detected [33]. Several of these proteins were also recently shown to be accessible to antibodies for *P. falciparum* sporozoites, and should now be considered during selection of antibody-based therapeutics and vaccines [6]. Lastly, the orthologues of 43 of 52 proteins (83%, Supporting Information Table 6) with known *P. falciparum* orthologues were also detected in our recent *P. falciparum* surface-exposed proteome, again lending support to the categorization of these proteins as being surface exposed.

Comparison of the PlasmoSEP predicted surface-exposed proteome with the experimentally determined surface-exposed proteome demonstrates the practical utility of our approach. Table 2 reports the predicted scores of PlasmoSEP, SignalP, and antigenicity predictors on a set of 37 identified high confidence (defined as having two or more unique peptides and/or published confirmation of surface exposure independent of mass spectrometric methods) SEPs on *P. yoelii*, ranked by the maximum score of the three predictors. Our approach confirms that 25 proteins are surface exposed with a prediction score ≥ 0.60. Interestingly, a careful examination of the 12 proteins not identified by our approach reveals that six of them are unlikely to be exposed to the surface and three of them are likely to be SEPs. This suggests that our approach is very promising in computationally assessing high-throughput results. Finally, we provide our predicted scores for the entire *P. yoelii* proteome in Supporting Information Table 7. Our predictions suggested that 159 proteins are expected to be surface exposed with prediction score ≥0.7 and 65 of these proteins have prediction scores ≥0.8.

It should be noted that the extremely small number (13) of known SEPs in our training data makes it very challenging to estimate a reasonable cut-off score (i.e., one that corresponds to a desired sensitivity-specificity tradeoff) for reliably

Table 2. List of 37 identified SEPs in *P. yoelii* salivary gland sporozoites using MS experiments and their predicted PlasmoSEP, SignalP, and antigenicity scores

ID	Name	PlasmoSEP	SignalP	Antigenicity	Max
PY17X_1330200	Glyceraldehyde-3-phosphate dehydrogenase, putative (GAPDH)	1.00	0.14	0.58	1.00
PY17X_0712100	Heat shock protein, putative (HSP70)	1.00	0.11	0.62	1.00
PY17X_1007600	Sporozoite invasion-associated protein 1 (SIAP1)	1.00	0.87	0.40	1.00
PY17X_1312400	Fructose-bisphosphate aldolase 2 (ALDO2)	1.00	0.10	0.31	1.00
PY17X_0420500	Alpha tubulin 1	1.00	0.14	0.45	1.00
PY17X_1354800	Sporozoite surface protein 2, thrombospondin-related anonymous protein (TRAP)	0.50	0.83	0.97	0.97
PY17X_1007700	Perforin-like protein 1,sporozoite micronemal protein essential for cell traversal (SPECT2)	0.80	0.64	0.56	0.80
PY17X_1461900	Actin I	0.80	0.10	0.52	0.80
PY17X_0835500	Conserved Plasmodium protein, unknown function	0.20	0.67	0.79	0.79
PY17X_0702200	Secreted ookinete protein, putative,GPI-anchored micronemal antigen, putative (GAMA)	0.00	0.76	0.48	0.76
PY17X_1427200	Conserved Plasmodium protein, unknown function	0.10	0.74	0.54	0.74
PY17X_0210500	Thrombospondin related sporozoite protein, putative (TRSP)	0.30	0.72	0.31	0.72
PY17X_1210100	Tubulin beta chain, putative	0.70	0.10	0.44	0.70
PY17X_0405400	Circumsporozoite (CS) protein (CSP)	0.70	0.68	0.70	0.70
PY17X_1037800	Glideosome associated protein with multiple membrane spans 3, putative (GAPM3)	0.70	0.12	0.14	0.70
PY17X_0902700.1	Merozoite adhesive erythrocytic binding protein (MAEBL)	0.30	0.69	0.44	0.69
PY17X_0826700	Phosphoglycerate kinase, putative (PGK)	0.00	0.10	0.68	0.68
PY17X_0912300	Conserved Plasmodium protein, unknown function	0.40	0.12	0.68	0.68
PY17X_0404800	Inner membrane complex protein 1a (IMC1a)	0.20	0.11	0.67	0.67
PY17X_1439800	Endoplasmin, putative (GRP94)	0.10	0.65	0.47	0.65
PY17X_1217500	Enolase, putative (ENO)	0.50	0.11	0.64	0.64
PY17X_1316500	Gamete egress and sporozoite traversal protein, putative (GEST)	0.20	0.63	0.46	0.63
PY17X_1034500	Rhoptry-associated protein 1, putative (RAP1)	0.00	0.62	0.22	0.62
PY17X_0910400	Carbonic anhydrase, putative	0.20	0.54	0.61	0.61
PY17X_1134900	Elongation factor 1-alpha, putative	0.60	0.12	0.45	0.60
PY17X_0703100	Protein disulfide isomerase, putative	0.10	0.59	0.49	0.59
PY17X_0404900	Membrane skeletal protein, putative	0.10	0.10	0.55	0.55
PY17X_0525300	Glideosome associated protein with multiple membrane spans 2, putative (GAPM2)	0.50	0.10	0.27	0.50
PY17X_1361400	Myosin A (MyoA)	0.20	0.10	0.40	0.40
PY17X_0303100	Hexose transporter (HT)	0.40	0.13	0.24	0.40
PY17X_0712800	14-3-3 Protein, putative (14-3-3I)	0.10	0.10	0.32	0.32
PY17X_0706500	Nucleoside transporter, putative (NT2)	0.20	0.11	0.32	0.32
PY17X_1424900	Conserved Plasmodium protein, unknown function	0.10	0.11	0.30	0.30
PY17X_0823700	Sugar transporter, putative	0.20	0.30	0.19	0.30
PY17X_0514100	Conserved Plasmodium protein, unknown function	0.00	0.10	0.22	0.22
PY17X_1143100	60S ribosomal protein L40/UBI, putative	0.00	0.12	0.09	0.12
PY17X_1118200	Histone H3 variant, putative (H3.3)	0.00	0.10	0.03	0.10

Our approach confirms that the first 25 proteins are SEPs with predicted score ≥0.60.

discriminating SEPs from non-SEPs. Estimation of such a cutoff score will have to wait until we accumulate a larger and diverse sample of known SEPs. Until then, our predictions should be viewed as a prioritized list of candidate SEPs for further experiments, which in turn can help improve the classifier, in an iterative fashion. Despite this limitation, we anticipate that these predictions will help to guide future experimental work for identifying novel SEPs in *P. yoelii*.

3.5 Application of PlasmoSEP to the human-infectious *P. vivax* malaria parasite

As *P. vivax* parasites cannot be continuously cultured in the laboratory, it is extremely difficult to conduct experimental determinations of the surface-exposed proteome of this malaria parasite species, even with access to patient isolates from endemic regions. To overcome this limitation, we have

instead turned solely to computational approaches. Buoyed by the success of the predictions of the PlasmoSEP algorithm with *P. falciparum* and *P. yoelii* proteomes, we have also applied this prediction software to the *P. vivax* proteome (Supporting Information Table 8). Several known SEPs (CSP, TRAP, IMC proteins, transporters, and proteins that are secreted during gliding and invasion) score positively using similar thresholds as were applied for *P. falciparum* and *P. yoelii*, indicating that the predictor accurately identifies SEPs. Moreover, as this is a proteome wide predictor, surface antigens from other life cycle stages (p25 and p28 are known surface proteins of the ookinete stage) also score positively, as this algorithm does not restrict the prediction only to sporozoites. This indicates that the same properties of these proteins may dictate their surface exposure throughout the life cycle. Lastly, a great number of the highest scoring proteins have not been experimentally or bioinformatically defined for *P. vivax*, and are currently noted as "hypothetical protein, conserved" in PlasmoDB. In light of this, our predictions that these are SEPs, which may be targetable by antibodies, may help to focus and prioritize future characterizations. Studies of these top candidates to first verify that they are indeed surface exposed, and then to also determine the importance of their biological function in the parasite will help guide efforts to rationally design subunit vaccines.

4 Concluding remarks

High-throughput MS-based proteomics has increasingly become the state-of-the-art experimental approach for mapping SEPs in many target pathogens [34, 35]. This is an essential and key step in developing novel subunit vaccines. However, due to technical limitations, the MS approach suffers from false positive as well as false-negative inferred protein identifications. We address this limitation by integrating high-throughput experimental studies, expert-annotated data, machine learning, and in silico bioinformatics tools to substantially improve the reliability of identification of SEPs in the malaria parasite. Our framework makes use of potentially labeled data (proteins tagged by an expert as surface or non-surface exposed) to build classifiers from small amount of labeled data as well as typically much larger amount of unlabeled data. By applying our approach to 205 experimentally determined SEPs of *P. falciparum* salivary gland sporozoite, we developed the PlasmoSEP classifier for predicting SEPs in *P. falciparum* from amino acid derived information. We used the PlasmoSEP classifier along with *in silico* bioinformatics tools for predicting secreted and protective proteins to filter out false positives from the *P. falciparum* SEPs identified using proteomics experiments, and to predict novel SEPs in *P. falciparum* proteome. To further assess the viability of PlasmoSEP, we used it to predict novel SEPs in *P. yoelii* (which were independently validated experimentally) and *P. vivax* malaria parasite.

The modularity of the PlasmoSEP framework allows it to be customized in several ways. For example, it can be easily modified to make use of potentially labeled data obtained from annotations supplied by multiple human experts or some in silico tools (e.g., tools for predicting protein subcellular localization). The framework can be used, in principle, to improve the reliability of the output of high-throughput experiments beyond the applications considered in this paper, as long as some labeled data, potentially labeled data, and unlabeled data are available. Work in progress aims to: (i) adapt other sophisticated semisupervised algorithms (e.g. [36, 37]) to learn predictive models from potentially labeled data; (ii) apply our framework to identify SEPs in other interesting pathogens, e.g., *B. pertussis* [38]; (iii) Develop a community resource for depositing the output of MS proteomics, enable community annotation and integration of data from multiple studies, and support the application of our framework to these data using Web browser based computational workflows.

We appreciate the assistance of Kristian Swearingen (ISB, CIDR) for the critical discussion of the data processing and analysis of surface proteomics data for Plasmodium parasites, and of Ben Allen (Penn State) for discussion of data analysis and interpretation. Experimental analyses were conducted at the Mass Spectrometry and Proteomics Resource Laboratory (MSPRL), Center for Systems Biology at Harvard University. This work was supported in part by the NIH under the NIAID Career Transition Award (1K22AI10139-01, S.E.L.), start-up funds provided by Penn State University (S.E.L.), the NACME Alfred P. Sloan Foundation Graduate Scholarship (E.E.M.), the ASM Robert D. Watkins Graduate Research Fellowship (E.E.M.) the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant UL1 TR000127 (V.H.), the Edward Frymoyer Endowed Professorship (V.H.), and the Center for Big Data Analytics and Discovery Informatics (V.H., Y.E.) which is cosponsored by the Institute for Cyberscience, the Huck Institutes of the Life Sciences, and the Social Science Research Institute at the Pennsylvania State University.

The authors have declared no conflict of interest.

5 References

[1] WHO, World malaria report 2014, WHO 2015. Available at: http://www.who.int/malaria/publications/world-malaria-report-2015/report/en/.

[2] Hoffman, S. L., Vekemans, J., Richie, T. L., Duffy, P. E., The march toward malaria vaccines. *Am. J. Prev. Med.* 2015, *49*, S319–S333.

[3] Neafsey, D. E., Juraska, M., Bedford, T., Benkeser, D. et al., Genetic diversity and protective efficacy of the RTS, S/AS01 malaria vaccine. *N. Engl. J. Med.* 2015, *373*, 2025–2037.

[4] Hodgson, S. H., Ewer, K. J., Bliss, C. M., Edwards, N. J. et al., Evaluation of the efficacy of ChAd63-MVA vectored vaccines expressing circumsporozoite protein and ME-TRAP against

controlled human malaria infection in malaria-naive individuals. *J. Infect. Dis.* 2015, *211*, 1076–1086.

[5] Lindner, S. E., Swearingen, K. E., Harupa, A., Vaughan, A. M. et al., Total and putative surface proteomics of malaria parasite salivary gland sporozoites. *Mol. Cell. Proteomics* 2013, *12*, 1127–1143.

[6] Swearingen, K. E., Lindner, S. E., Shi, L., Harupa, A. et al., Interrogating the plasmodium sporozoite surface: identification of surface-exposed proteins and demonstration of glycosylation on CSP and TRAP by mass spectrometry-based proteomics. *PLoS Pathog.* 2016, *12*, e1005606.

[7] Kotsiantis, S., Supervised machine learning: a review of classification techniques. *Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*. IOS Press, Netherlands 2007, pp. 3–24.

[8] Drabovich, A. P., Pavlou, M. P., Batruch, I., Diamandis, E. P., Proteomic and mass spectrometry technologies for biomarker discovery, in: Haleem, J. I. and Timothy, D. V. (Eds.), *Proteomic and Metabolomic Approaches to Biomarker Discovery*, Elsevier, Netherlands 2013, pp. 17–37.

[9] Bock, T., Bausch-Fluck, D., Hofmann, A., Wollscheid, B., CD proteome and beyond–technologies for targeting the immune cell surfaceome. *Front. Biosci.* 2012, *17*, 1599–1612.

[10] Zhu, X., Goldberg, A. B., Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning* 2009, *3*, 1–130.

[11] Chapelle, O., Schölkopf, B., Zien, A., *Semi-Supervised Learning*, MIT Press, London, England 2006.

[12] Fischer, B., Grossmann, J., Roth, V., Gruissem, W. et al., Semisupervised LC/MS alignment for differential proteomics. *Bioinformatics* 2006, *22*, e132–e140.

[13] Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O., Borodovsky, M., Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* 2005, *33*, 6494–6506.

[14] Ou, H.-Y., Guo, F.-B., Zhang, C.-T., GS-Finder: a program to find bacterial gene start sites with a self-training method. *Int. J. Biochem. Cell Biol.* 2004, *36*, 535–544.

[15] Stanescu, A., Caragea, D., Semi-supervised self-training approaches for imbalanced splice site datasets. *The Sixth International Conference on Bioinformatics and Computational Biology (BICoB)*, Las Vegas, NV 2014, pp. 131–136.

[16] Tang, S., Lomsadze, A., Borodovsky, M., Identification of protein coding regions in RNA transcripts. *Nucleic Acid Res.* gkv227, 2015.

[17] Xu, Y.-Y., Yang, F., Shen, H.-B., Incorporating organelle correlations into semi-supervised learning for protein subcellular localization prediction. *Bioinformatics* btw219, 2016.

[18] El-Manzalawy, Y., Honavar, V., A framework for developing epitope prediction tools. *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology*, ACM, Niagara Falls, NY 2010, pp. 660–662.

[19] Hall, M., Frank, E., Holmes, G., Pfahringer, B. et al., The WEKA data mining software: an update. *ACM SIGKDD Explor. Newslett.* 2009, *11*, 10–18.

[20] Petersen, T. N., Brunak, S., von Heijne, G., Nielsen, H., SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* 2011, *8*, 785–786.

[21] Blobel, G., Dobberstein, B., Transfer of proteins across membranes. I. Presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma. *J. Cell Biol.* 1975, *67*, 835–851.

[22] Coleman, J., Inukai, M., Inouye, M., Dual functions of the signal peptide in protein transfer across the membrane. *Cell* 1985, *43*, 351–360.

[23] Bradley, A. P., The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* 1997, *30*, 1145–1159.

[24] Bishop, C. M., Training with noise is equivalent to Tikhonov regularization. *Neural Comput.* 1995, *7*, 108–116.

[25] Qi, Y., Random forest for bioinformatics, in: Zhang C., Ma Y. (Eds.), *Ensemble Machine Learning*, Springer, New York 2012, pp. 307–323.

[26] Cha, S.-J., Kim, M.-S., Pandey, A., Jacobs-Lorena, M., Identification of GAPDH on the surface of Plasmodium sporozoites as a new candidate for targeting malaria liver invasion. *J. Exp. Med.* 2016, *213*, 2099–2112.

[27] Rappuoli, R., Reverse vaccinology, a genome-based approach to vaccine development. *Vaccine* 2001, *19*, 2688–2691.

[28] Donati, C., Rappuoli, R., Reverse vaccinology in the 21st century: improvements over the original design. *Ann. N. Y. Acad. Sci.* 2013, *1285*, 115–132.

[29] Ortiz, D., Guiguemde, W. A., Johnson, A., Elya, C. et al., Identification of selective inhibitors of the Plasmodium falciparum hexose transporter PfHT by screening focused libraries of anti-malarial compounds. *PLoS One* 2015, *10*, e0123598.

[30] Arumugam, T. U., Takeo, S., Yamasaki, T., Thonkukiatkul, A. et al., Discovery of GAMA, a Plasmodium falciparum merozoite micronemal protein, as a novel blood-stage vaccine candidate antigen. *Infect. Immun.* 2011, *79*, 4523–4532.

[31] Engelmann, S., Silvie, O., Matuschewski, K., Disruption of Plasmodium sporozoite transmission by depletion of sporozoite invasion-associated protein 1. *Eukaryot. Cell* 2009, *8*, 640–648.

[32] Carey, A. F., Singer, M., Bargieri, D., Thiberge, S. et al., Calcium dynamics of Plasmodium berghei sporozoite motility. *Cell. Microbiol.* 2014, *16*, 768–783.

[33] Heintzelman, M. B., Gliding motility in apicomplexan parasites. *Seminars in Cell and Developmental Biology*, Elsevier, Netherlands 2015, pp. 135–142.

[34] Cordwell, S. J., Technologies for bacterial surface proteomics. *Curr. Opin. Microbiol.* 2006, *9*, 320–329.

[35] Olaya-Abril, A., Jiménez-Munguía, I., Gómez-Gascón, L., Rodríguez-Ortega, M. J., Surfomics: shaving live organisms for a fast proteomic identification of surface proteins. *J. Proteomics* 2014, *97*, 164–176.

[36] Criminisi, A., Shotton, J., Konukoglu, E., Decision forests: a unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends® in Computer Graphics and Vision 7*, 2012, pp. 81–227.

[37] Bennett, K., Demiriz, A., Semi-supervised support vector machines, in: Kearns, M. S., Solla, S. A., Cohn, D. A. (Eds.), *Advances in Neural Information Processing Systems*, The MIT Press, Cambridge, MA 1999.

[38] Raeven, R. H., van der Maas, L., Tilstra, W., Uittenbogaard, J. P. et al., Immunoproteomic profiling of Bordetella pertussis outer membrane vesicle vaccine reveals broad and balanced humoral immunogenicity. *J. Proteome Res.* 2015, *14*, 2929–2942.

